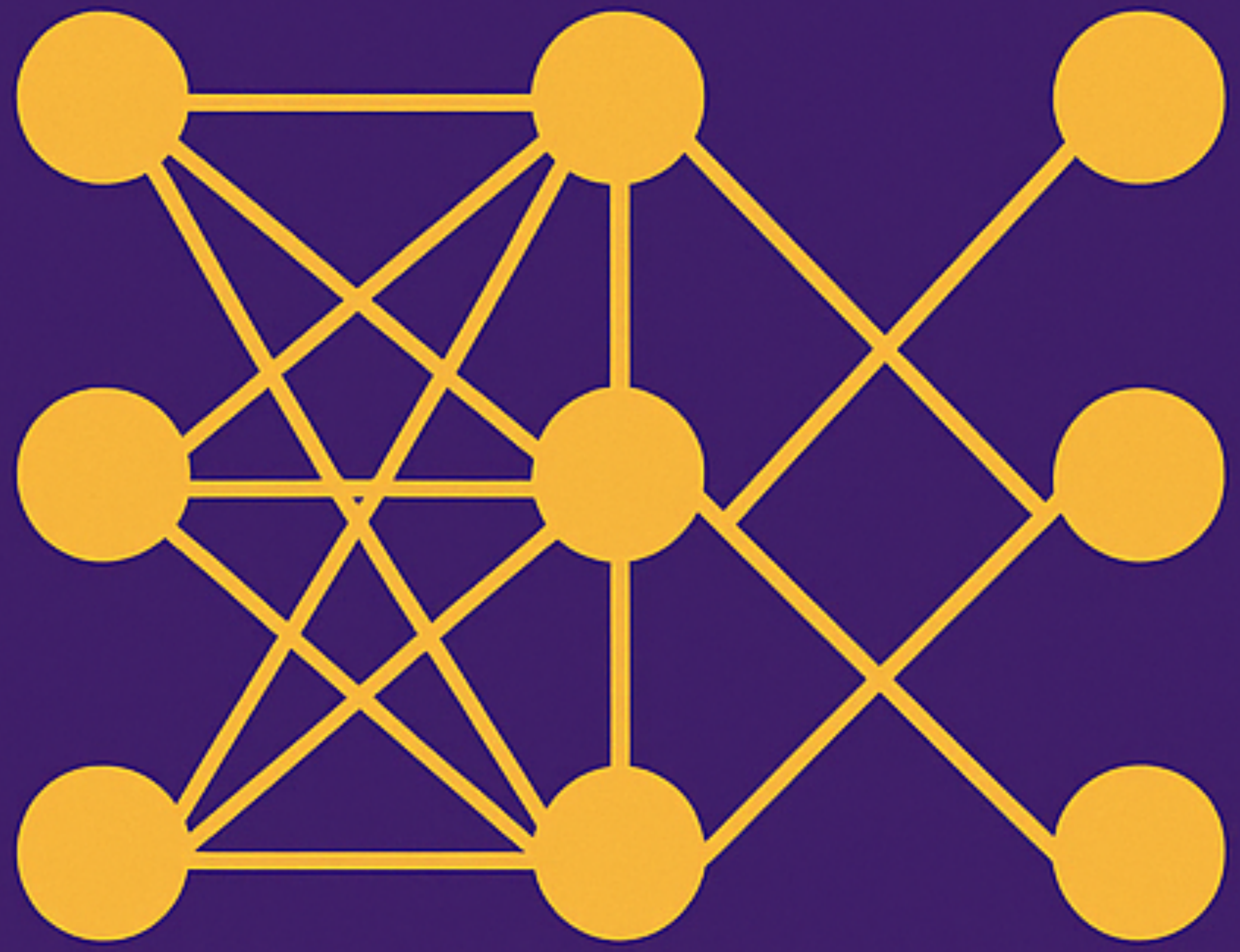
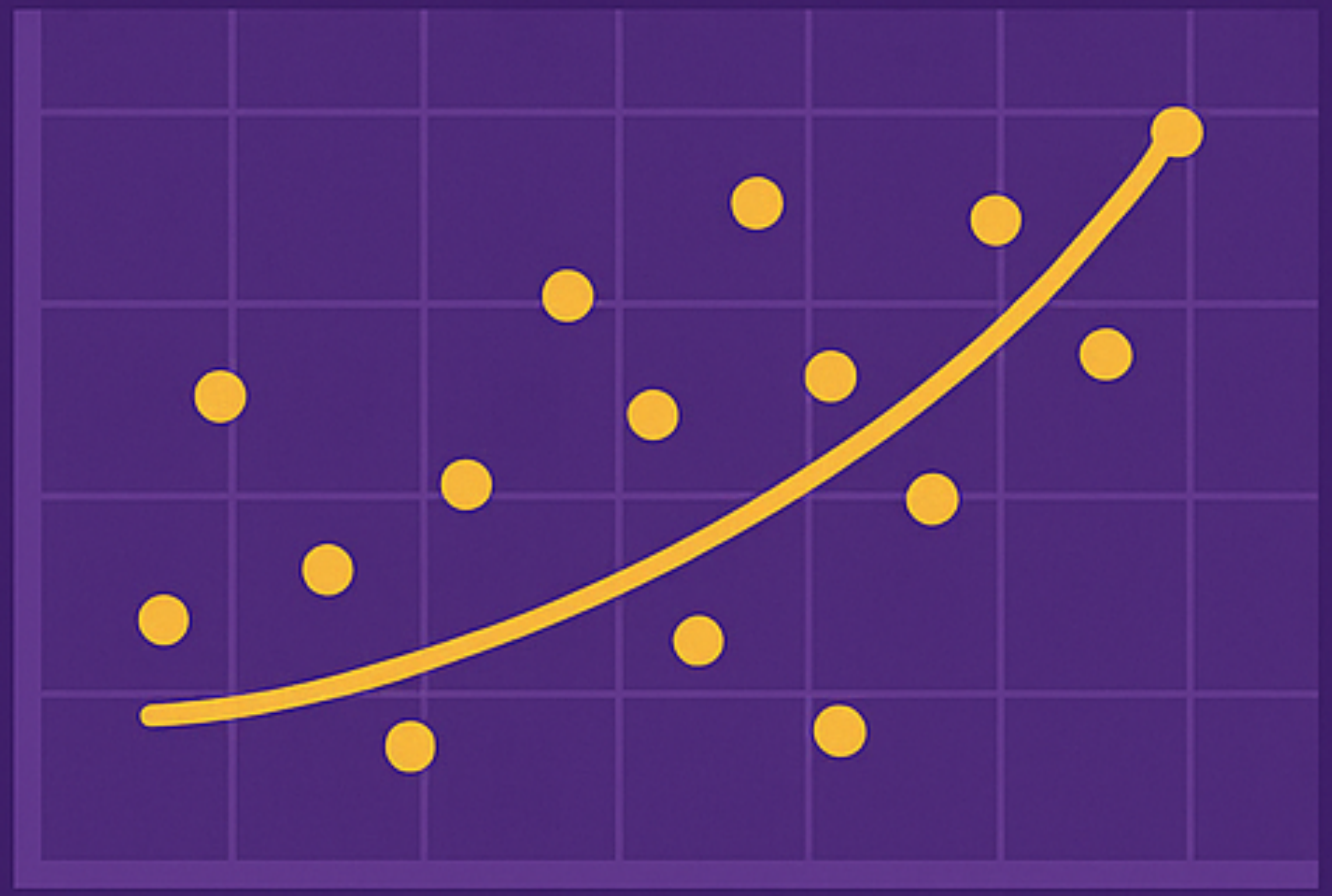
Hot to write about ML

On a Physics/Astrophysics paper

INTRODUCTION TO MACHINE LEARNING

FALL 2025 LSU





Use of Machine Learning

- **Motivation**: is the use of ML well justified?
  - Avoid to use ML just to use ML

- **Feasibility**: is the the dataset good enough?
  - Garbage in —> Garbage out
  - Too small of a dataset

Describe the Dataset Clearly

1. Origin and justification of the data
2. Preprocessing steps (normalization, filtering, cleaning)
3. Handling of missing values
4. Train/validation/test splits
5. Data augmentation (if used)

Document the Model Architecture

1. Type of model (e.g., CNN, RNN, transformer, XGBoost)
    - ✴ If famous published architecture (e.g. ResNet18), reference the paper!
    - ✴ if Custom NN, report the details + graphic scheme
2. Architecture details: number of layers, number of channels, units, activation functions
3. Explanation of why the chosen architecture is appropriate for the physics problem

Report Hyperparameters and Training Procedure

1. Learning rate, batch size, number of epochs
2. Optimizer (Adam, SGD, RMSProp, etc.)
   - ✳ Reference the papers!
3. Regularization methods (dropout, weight decay, early stopping)
4. Loss function used and justification (e.g., MSE, cross-entropy)
   - ✳ Show the Train+Validation Loss curve (appendix is fine)
5. Hardware/software environment
6. This ensures the experiment is fully reproducible.

Include Robust Model Evaluation

1. Performance metrics (accuracy, F1, ROC curve, MSE, etc.)
2. Comparison to baselines (e.g., simple physical model, traditional classifier)
3. Cross-validation for limited datasets
4. Calibration and uncertainty quantification (especially in physics) — Read the literature, google it, ask an expert (consult a data scientist, machine learning expert)
5. Use IAI/XAI Tools to Interpret Results
   - ✴ Identify which features are important
   - ✴ Visualize saliency or attention maps
   - ✴ Check for physically meaningful behavior (e.g., monotonicity where expected)

Ensure Full Reproducibility (as much as possible)

1. Provide code and trained models (or at least scripts)

2. Provide full details of data processing and random seeds

3. Document software versions

Avoid Overclaiming

2. Do not imply causation from a purely correlational ML model
3. Acknowledge model limitations and uncertainties
4. Discuss potential dataset biases and how they may affect conclusions